カラムデータベースにおける理解支援 : 展望と周辺技術

AUTHOR(S):

亀田, 尭宙

CITATION:

亀田, 尭宙. カラムデータベースにおける理解支援 : 展望と周辺技術. CIAS discussion paper No.53 : 「カラム」の時代 VI.--近代マレー・ムスリムの日常生活2 2015, 53: 10-13

ISSUE DATE:

2015-03

URL:

http://hdl.handle.net/2433/228636

カラムデータベースにおける理解支援
展望と周辺技術

亀田尭宙

はじめに

京都大学地域研究統合情報センター（以下、地域研）では、研究者の研究資源を共有すべく、Myデータベースシステムが構築されている。CSVのような表形式のデータをアップロードしデータベース構造に関する情報を設定するだけで、データの検索、APIの提供を可能にしている。カラムデータベースも、データはMyデータベースに入っており、システムが提供する元々の検索インタフェース[1)]も備えている。一方で、データの構造や内容を知らない利用者は検索ワードを思い浮かべることができない。これを解決するために、ワードクラウド（図１）による検索や、号ごとに表紙を並べた一覧ページ用意することにより、より幅広い人に使ってもらうことができるようなサイトも、MyデータベースのAPIを利用して構築されている[2)]。また、JawiとRumi[3)]を左右に並べて閲覧できるインタフェースなども工夫されている［山本 2014］。

つまり、カラムのデータ化、ローマ字翻字、検索・閲覧のインタフェースという土台は既にできあがったと言える。次の方向性の一つは、カラム以外の雑誌の収集とデータベース化であり、雑誌の枠を越えた引用や論争の多いマレー・インドネシア語定期刊行物の言論空間において、それを一体として捉えることは非常に重要である。

一方で、言論空間を一体として捉えるためには、言及先の同定や、同じトピックを扱っている記事の関連付け、議論におけるロジックやレトリックの理解が必要になってくる。本稿では、カラムについて利用者が理解を深めるためにそれらを情報技術で支援するにあたって、展望について周辺技術を含めつつ紹介する。大きく分けると２つの支援を想定している。１つ目は、Entity Linking と呼ばれる技術で、カラムの文章内に現れる専門用語や引用記述などについて、対応する外部データベースのデータを見つける技術である。２つ目は、レトリックの研究である。言い換えると、「何が」語られているか、それが「どう」語られているか、という２つの側面で整理することで、カラムの理解を支援したいと考えている。

図１　質問コーナーでよく使われる単語を検索クエリとして提示しているワードクラウド

Entity Linking

カラムの中には、クルアーンをはじめとして、他の書物、雑誌やニュースに対する言及が多数出現する。これは、カラムの時代を読み解くにあたって、カラムだけを見ていては不十分で、周りの資料やそれら資料との関係を読み解いていくことが必要だということを意味している。そして、その読み解きの一歩として、言及それぞれの言及先を同定することは理解の助けになるだろう。

1）http://app.cias.kyoto-u.ac.jp/infolib/meta_pub/G0000003qalam2
2）http://majalahqalam.kyoto.jp/
3）アラビア文字で書かれたマレー語はJawi、ローマ字で書かれたマレー語はRumiと呼ばれる

表1　カラムに出てくる引用文(左)とクルアーンデータベース内のマレー語翻訳文(右)	
Sesungguhnya Allah tidak akan mengampunkan yang ia disekutukan, tetapi ia akan mengampuni selain daripada itu bagi sesiapa yang ia kehendaki: dan barangsiapa yang menyekutukan Allah maka sesungguhnya ia telah membuat suatu dosa yang terang dan nyata	Sesungguhnya Allah tidak akan mengampunkan dosa syirik mempersekutukanNya (dengan sesuatu apajua), dan akan mengampunkan dosa yang lain dari itu bagi sesiapa yang dikehendakiNya (menurut aturan SyariatNya). Dan sesiapa yang mempersekutukan Allah (dengan sesuatu yang lain), maka sesungguhnya ia telah melakukan dosa yang besar.

クルアーンの引用については、例えば、カラムの「Bidasan Kepada Faham Tak Bertuhan」[Qalam 1950.12:11]の記事において、「Sesungguhnya Allah tidak akan mengampunkan yang ia disekutukan …」といった章句が引用されている。これは、4章48節 の文言に対応しているが、quran.com[4] のようなデータベース内の章句と一字一句対応しているわけではないが(表1)、コンピュータによる対応の発見が可能だろう[ブルドン宮本 & 山本 2013]。

資料間の言及については、たとえば [光成 2012]では、カラムの「Kahwin Paksa dan Wali Mujbir(強制婚と強制後見)」[Qalam 1954.4: 29]の記事において、ジャカルタの新聞『アバディ』から記事を転載している例が示されている。こちらの場合は『アバディ』の方がデータベース化されていないため、自動で対応付けることはできない。

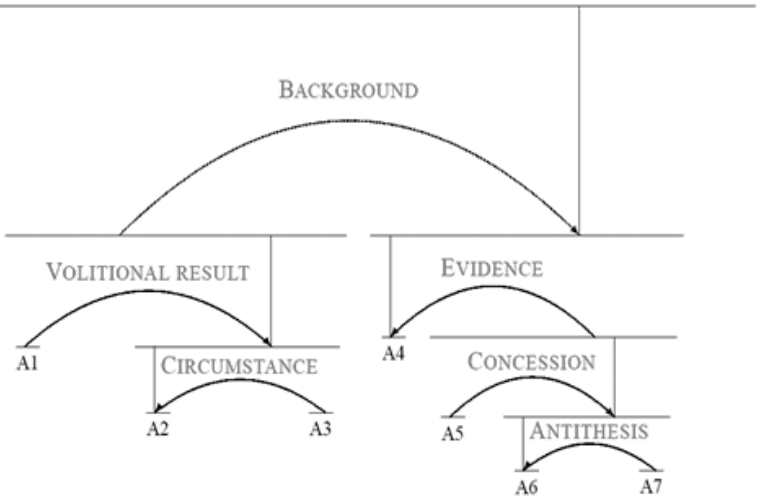
ここまでは文レベルでの関連付けを見てきたが、語レベルでの関連付けも有用だと考えられる。例えばIslah や Kafir といった言葉は、イスラムの専門用語で、それぞれ reform や unbeliever と英訳されることが多いものの、その訳のみでは意味を捉えきれず、理解には背景知識が必要になる。その場合は、適切な解説にリンクすることで利用者の理解を促進することが可能になる。

言及先の同定には複数の関連技術、関連研究が存在する。人名、地名など特定のタイプの用語のことをNamed Entity と呼び、文書内からそういった用語を発見することを Named Entity Recognition (以下NER)、それを外部データベースと関連付けることをNamed Entity Linking (以下NEL)と呼ぶ。良くつかわれる外部データベースの一つに Wikipedia[5]があり、Wikipediaのデータと結びつけることは俗に Wikifyとも呼ばれる。これら単語やフレーズレベルの同定技術は、引用文のような文章レベルで用いることもできる。用語や文章とその文脈を手掛かりに類似性を測り対応するデータを探すという技術的枠組みは変わらない。

Entity Linking の手法は [Shen 2014] に最新のサーベイがある。多くの研究で、対応する候補の選定、類似性による候補のランキング、対応先が無いことの検知の3つのモジュールによってEntity Linkingが行われているとまとめられている。また、対応付け先のデータベースとしてはWikipediaの他、DBpediaやYAGOといった汎用的なデータベースが代表的なものとして挙げられているが、個々の分野に特化したEntity Linkingの研究もある。例えば、絶滅危惧種の種名からデータベース内にある種の情報へリンクする研究を筆者は行っている [Kameda et al. 2013]。類似性による候補のランキングについてはレーベンシュタイン距離を測るといった表層的なマッチングから、Latent Dirichlet Allocation のように語の背後にあるトピックを推定した上でマッチングを考えるものまでさまざまであるが、引用のように文言をそのまま使いまわすことが多い場合は表層的なマッチングが有効と考えられる[6]。ただし、マレー語は接頭辞や接尾辞による変化が豊富であり、それらを省くような前処理を行うか、Longest Common Subsequenceのようにそういった変化と関わりなく類似性を測れる指標を使うように留意せねばならない。また、多くの論文で実験データセットとして用いられているのは、NIST Text Analysis Conference(TAC)[7] のKnowledge Base Population (KBP)タスクのデータであるが、これは英語に限られている。言語横断でのEntity Linking [Mcnamee et al. 2011] では、対訳対のあるコーパスの英語側のみにNamed Entity のアノ

4) The Noble Qur'an http://quran.com/
5) Wikipedia https://ja.wikipedia.org/ (日本語版へのリンク)
6) 引用ではないが、同じく語の使いまわしが多い、論文の概要と本文における対応発見において、表層的なマッチングが比較的機能することを『論文における要約記述に対応するパラグラフの同定手法』(亀田尭宙 et al., 2013. 人工知能学会全国大会(第27回)論文集)で発表しており、今回も表層的な手がかりを中心に対応付けを行いたいと考えている。
7) Text Analysis Conference (TAC) Overview http://www.nist.gov/tac/about/index.html

テーションを行って、Berkley Aligner[8]を適用し対応する語を特定することで、独自にデータセットを作っている[9]。英語以外の言語でのコーパスやデータセットは自然言語処理研究において不足しており、カラムデータベースを元に、このようなタスクの評価に用いることができるデータセットを作ることで、他の研究者との協働も促進したいと考えている。



[Farmington police had to help control traffic recently]^{A1} [when hundreds of people lined up to be among the first applying for jobs at the yet-to-open Marriott Hotel.]^{A2} [The hotel's help-wanted announcement – for 300 openings – was a rare opportunity for many unemployed.]^{A3} [The people waiting in line carried a message, a refutation, of claims that the jobless could be employed if only they showed enough moxie.]^{A4} [Every rule has exceptions,]^{A5} [but the tragic and too-common tableaux of hundreds or even thousands of people snake-lining up for any task with a paycheck illustrates a lack of jobs.]^{A6} [not laziness.]^{A7}

(*The Hartford Courant*, editorial)

図3　[Mann & Thompson 1988]における文の解析例

レトリック

カラムの中には、1001 Masalah（千一問）と題された質問コーナーがあり、そこでは日常的な読者の疑問にイスラムの立場から回答がなされている。例えば、Q「バドミントン、ホッケー、サッカーをすることはイスラムの教えに反するか？」A「ゲームはイスラムの教えに反しない。それはレクリエーションだ。しかし、男性と女性で混ざって行った場合は教えに反する。」といった問答が見られる［*Qalam* 1952.1: 35］（図２）。内容に着目すると、価値観や道徳観念の変遷を辿ることができるが［金子 2014］、一方で、答え方のレトリックも興味深い。先に挙げたように端的に答えたのちに条件を付すものも多くみられる一方、国による産児制限の考えに従ってよいかという質問に対しては、「この質問に答える前に、産児制限や家族計画の目的について知る必要がある。」と前置きして、国家政策や家計の現実的な説明を踏まえ「子供の数を抑えることが求められている」と理解を示したのち、ただ一文「イスラムの教義に従えば、子どもは神からの授かりものであり、それを人間がどうにかすることはできない。」と述べることで直接的に回答しない形がとられている［Qalam 1951.12: 41］。

----------✱----------✱----------

Hussin, Parit Lat, Nombor 1 Jalan Baharu, Parit Buntar Perak bertanya: Adakah permainan badminton, hoki dan football, itu haram?

Jawab : Permainan-permainan itu tidak haram. Ia suatu riadah. Tetapi yang haram di dalamnya ialah aurat yang terbuka demikian juga pergaulan di antara lelaki dengan perempuan.

----------✱----------✱----------

図2　1001 Masalah の問答例

レトリックの研究は古くからおこなわれており、Rhetorical Structure 理論［Mann & Thompson 1988］は要約の生成や文章の生成および解析の手法として用いられてきた。これは文の間の関係に着目して、ツリー上に文章の構造を解析する理論である（図３[10]）。近年ではその分析をコンピュータによって行う手法が開発されており、表層的な手がかりを使ったもの［Marcu 1997］から一階述語論理の構造を手かったもの［Subba 2007］までさまざまある。よりマクロな構造の分析としては背景説明や対照比較といったレトリックの機能を共有している文のまとまりによって科学論文の構造を分析する Argumentative Zoning 理論がある［Teufel 2010］。こちらもコンピュータによる分析の手法が洗練されてきており、例えばトピックから単語を生成するモデルとレトリックから単語

8）berkeleyaligner - A word alignment software package for machine translation - Berkley Aligner https://code.google.com/p/berkeleyaligner/

9）Cross-Language Entity Linking http://pmcnamee.net/xlel.html に公開されている。

10）図自体は分かり易く描かれていた[Forsbom 2005]のものを掲載している。

を生成するモデルを合わせることでArgumentative Zoneごとの単語の分布をモデリングする手法が提唱されている[Ó Séaghdha & Teufel 2014]。

おわりに

カラムで何がどう語られているかの理解を支援するための技術について調査し展望を述べた。今後は具体的な事例を作って、有効性を検証しながら理解支援の技術を実装していきたいと考えている。

参考文献

Forsbom, E., 2005. Rhetorical Structure Theory in Natural Language Generation.

Kameda, A. et al., 2013. Integrate Japanese Red List into LOD of Species. P*NC Annual Conference and Joint Meetings 2013*.

Mann, W.C. & Thompson, S. a, 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Natural Language Generation: New Results in Artificial Intelligence, Psychology and Linguistics*, 8, pp.243–281.

Marcu, D., 1997. The rhetorical parsing of natural language texts. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pp.96–103.

Mcnamee, P. et al., 2011. Cross-Language Entity Linking. *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP-2011)*, pp.255–263.

Ó Séaghdha, D. & Teufel, S., 2014. Unsupervised learning of rhetorical structure with un-topic models. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp.2–13.

Shen, W., 2014. Entity Linking with a Knowledge Base : Issues Techniques, and Solutions. *Knowledge and Data Engineering, IEEE Transactions*, 27(2), pp.1–20.

Subba, R., 2007. Exploiting Event Semantics to Parse the Rhetorical Structure of Natural Language Text. *Proceedings of the NAACL-HLT 2007 Doctoral Consortium*, pp.21–24.

Teufel, S., 2010. *The Structure of Scientific Articles: Application to Citation Indexing and Summarization*, Stanford, CA: CSLI Publications.

ブルドン宮本ジュリアン & 山本博之 2013「アラビア文字・多言語文書の横断検索システム構築――『カラム』記事のコーラン引用部分表示の試み」『『カラム』の時代Ⅳ――マレー・ムスリムによる言論空間の形成』pp. 9–20.

金子奈央, 2014.「マレー・コミュニティにおける家族・子ども・教育」『『カラム』の時代Ⅴ――近代マレー・ムスリムの日常生活(CIAS Discussion Paper No. 40)』pp. 24–28.

光成歩 2012「1950年代の『強制婚』論議にみるカラム誌の改革論理」『『カラム』の時代Ⅲ――マレー・イスラム世界におけるイスラム的社会制度の設計(CIAS Discussion Paper No. 23)』pp. 40–47.

山本博之 2014「東南アジアの現地語文献のデジタル・アーカイブ化プロジェクト――2013年度の活動紹介」『『カラム』の時代Ⅴ――近代マレー・ムスリムの日常生活(CIAS Discussion Paper No. 40)』pp. 35–41.